iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50373

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Analysis of Hybrid Machine Learning Algorithm for Heart Disease Prediction

Mihir Walvekar[1], Shivam Gupta[2], Mayur Bhoire[3], Sohail Khan[4], Alpana Borse[5]
[1, 2, 3, 4,5]*Information Technology Department, Pimpri Chinchwad College of Engineering, Pune*

Abstract: ***In the current environment of rapid technological and data advancements, the healthcare sector is one of the most exciting fields of research. It's challenging to deal with a lot of patient data. Big data analytics makes it simple to process this data. Around the world, numerous treatments exist for various diseases. A novel method that aids in disease prediction in machine learning. In this study, machine learning is used to forecast diseases based on symptoms. The presented data set is used to train machine learning algorithms like Adaboost, XGBoost, CNN, Naive Bayes, Decision Trees, and Random Forests to predict disease. Research reveals the most accurate algorithm. Performance on a specific dataset is what determines an algorithm's accuracy. Then We Justify that The Person Having Heart Diseases or Not. With Rapid Advancement in Technology, our model is trained so well that we can predict heart diseases early.***

Keywords: AdaBoost, Convolutional Neural Network (CNN), Decision Tree (DT), K-Nearest Neighbor (KNN), Machine Learning (ML), Naïve Bias, Random Forest (RF), XGboost.

I. INTRODUCTION

Most people—50% of the population—have one or more chronic diseases, and as a result, these individuals spend more money on medical care [1]. The incidence of sickness is rising as lifestyles are becoming better. Nearly 61% of deaths in India are caused by non-communicable diseases like diabetes, cancer, and heart conditions. Diseases are mostly brought on by environmental factors and human lifestyle choices. IOT-based illness prediction is utilised to get early disease detection, lower disease risk, and disease diagnosis [3]. The biggest issue with IOT-based illness prediction, however, is that patients must wear more gadgets, which is uncomfortable and difficult for them to do [3].

Data mining has therefore become a preferred approach in the healthcare industry today for the purposes of illness prediction, detection, and diagnosis. Data mining is the process of obtaining the necessary information from a sizable historical data set. Future predictions are made using historical facts from the past [8]. Data mining encompasses a wide range of disciplines, including statistics, artificial intelligence, database design, and machine learning [8]. Similar to this, data mining is used in the medical industry to uncover hidden patterns [8].

Making decisions gets challenging since most medical data has concealed information. In order to analyse data and uncover hidden patterns in medical records, machine learning is crucial. Machine learning is used in a variety of industries, including banking, government, transportation, healthcare, and marketing [6]. Data mining has a subset called machine learning that deals with a lot of well-organized data.

Machine learning is employed in the medical industry for illness prediction, detection, and diagnosis [6]. The main objective of these procedures is to detect heart disease sooner so that it can lead to a quicker diagnosis and better heart disease therapy. In previous years, a lot of features from structural data were automatically extracted for use in medical research. Three different categories of data—structured, unstructured, and semi-structured—are included in the database.

People now have to visit a doctor when they contract certain ailments, which is costly and time-consuming. The fact that the condition cannot be diagnosed even when the user is far from a doctor or hospital can be challenging for the user. As a result, if the aforementioned procedure can be carried out by a computer program that saves time and money, it might be simpler for patients who can facilitate the process.

Data mining techniques are used in other systems for prediction of Heart Disease. With help of user-specified symptoms, Heart Disease Predictor web application may forecast a user's risk of developing Heart Disease. A dataset compiled from several health-related websites is used to forecast Heart Disease. Users of a Heart Disease predictor can learn the likelihood of diseases with particular symptoms. When a person has a certain disease, they don't immediately have an option. As a result, this method helps patients to predict Heart Disease around the clock.

II. PROPOSED METHODOLOGY

This paper's main goal is to determine whether the patient has heart disease or not. Determining if a patient's risk of developing heart disease is high or low. The user inputs the necessary numbers from his or her health report. The historical dataset is then uploaded. The majority of the datasets in the medical industry may include missing values, making accurate prediction challenging. Therefore, imputation and data cleaning steps are required for this missing data. After this data imputation, we must use a data cleaning and data imputation method to convert the missing data into structured data. Then, using the input values as a basis, the naive bayes and KNN algorithm is used to forecast heart disease. The first of the four algorithms we'll look at is naive bayes and KNN for classification. However, the classifier value that had the highest accuracy was used as input to the Random Forest algorithm for risk prediction. Since the performance of the naive bayes classifier is superior to that of the other two classifiers, it provides input to the random forest. We are able to determine the patient's risk level by applying the Random Forest method. For feature extraction, the Convolutional neural network approach is employed. And based on those characteristics, the softmax classifier is used to determine the risk classification for heart disease.

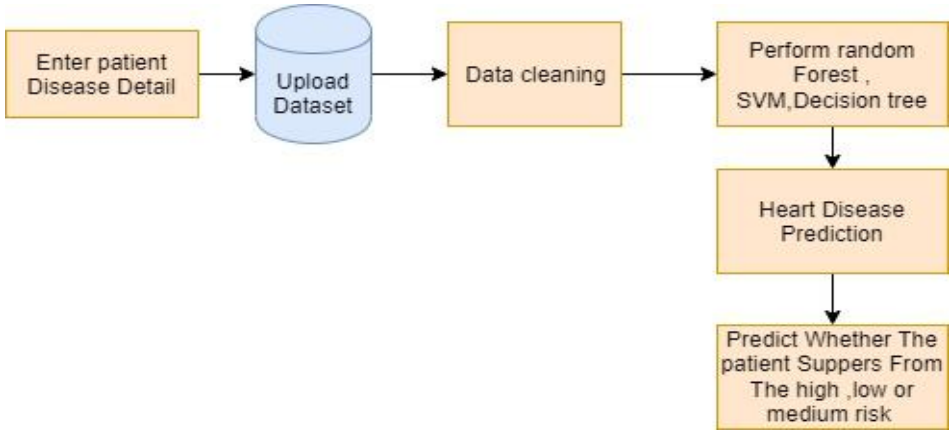


Fig. Proposed System Architecture

A. Dataset

We are leveraging data from the UCI Repository for the prediction of heart disease and heart risk. The dataset includes 12 attributes, including age, sex, the type of chest pain, blood pressure, cholesterol, the number of major blood vessels coloured by fluorescence, fast blood sugar, the ST segments that are induced by exercise relative to rest, the peak exercise ST segments, defect values, and exercise-induced angina, among others. The real heart disease prognosis was made using these characteristics. The risk of heart disease is calculated using the results of both the KNN and the Naive Bayes algorithms, but in this case, we primarily use the results of the Naive Bayes algorithm to estimate the risk in percentage terms and to show whether it is high, low, or medium.

B. Data Cleaning and Data Imputation

Unstructured data, or data that is not in a well-formed format, makes up a dataset. Most medical data is not formatted correctly. Data cleaning and data imputation are required for the missing data. To obtain organised data, the dataset must be cleaned up of undesired and noisy data.

C. Disease Risk Prediction

The primary goal of this article is to use the heart disease dataset to predict whether a patient is at high or low risk for developing heart disease. We use a deep learning algorithm to create the risk prediction model for heart disease, using inputs like age, sex, blood pressure, and others. The output value, h, indicates whether the patient is at high or low risk for developing heart disease. Specifically, h=h1, h2, where h1 indicates a high risk for developing heart disease and h2 indicates a low risk.

D. Evaluation Methods

Following are the notations we have for the experimental result:
TN: True Negative (correctly predicted the number of instances as not required),
FP: False Negative (incorrectly predicted the number of instances as not required),

TP: True Positive (number of instances correctly predicted),
FP: False Positive (number of instances incorrectly predicted)
Four measurements can be calculated based on this characteristic.

1) ACCURACY is equal to TP+TNTP+FP+TN+FN
2) PRICE = TP TP+ FP
3) Recall = TP+TP+FN
4) F1-Measure = 2 (Precision * Recall) / (Recall + Precision)

Users can publish their queries on the system to receive additional system benefits, which will let patients know about any responses from other doctors or specialists on their condition. This system's benefit is that it informs users of the risk of developing heart disease.

III. ALGORITHMS

We are employing a deep learning system to forecast the risk of sickness. Convolutional neural networks, a type of deep learning technique, automatically extract text features. Here, the CNN-UDRP algorithm is being used to forecast the disease's risk. Prior to that, it uses the Nave Bayes and KNN algorithms to forecast cardiac disease.

A. Naïve Bayes

Naïve bayes classifier based on probabilistic model and depends on the bayes theorem. In the supervised learning, the naïve bayes classifier work. The particular features which is describe in a class that are not related to the another features.

P(c/y) = P(y/c)* P(C)/ P(y)
P(c/y) = posterior probability,
P(c) = prior probability of class,
P(y/c) = likelihood probability of the class,
P(y) = prior probability of predictor.
On the bases of this algorithm, the classification is carried out.

B. KNN Algorithm

KNN is a classifier that stores all of the variable values, using those records as a foundation to classify the variable's unknown value. The variable's similarity index includes the unknown value. A non-parametric classification technique is KNN. The first type of KNN is a structure-less NN approach, whereas the second type is a structure-based NN technique. There are training and testing data for the structured based NN in that set of data.

C. Random Forest

A well-liked machine learning technique called Random Forest is used for classification, regression, and other applications. Multiple decision trees are used in this ensemble learning technique to provide predictions. A huge number of decision trees are constructed in a Random Forest using a randomly chosen subset of the data's characteristics. The ultimate forecast is the result of the majority vote of all the trees in the forest, each of which independently offers a prediction. This strategy lowers the possibility of overfitting and increases the model's precision.

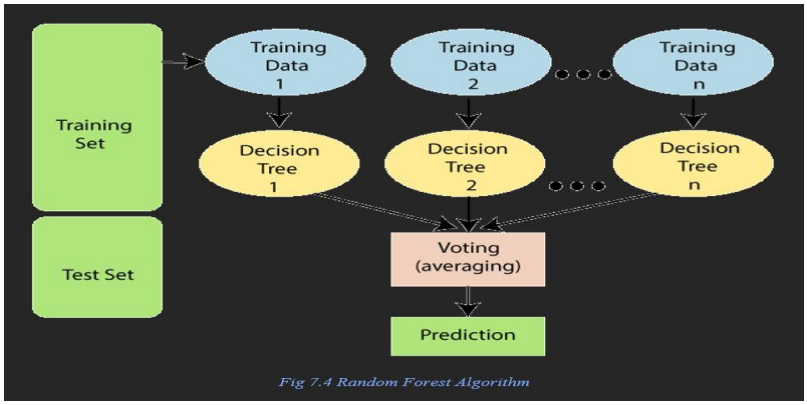


Fig 7.4 Random Forest Algorithm

D. Decision Tree

A well-liked machine learning technique called Decision Tree is utilised for both classification and regression problems. The programme builds a model of decisions and potential outcomes that resembles a tree. Each leaf node in the tree represents a class label or a numerical value, whereas each interior node represents a decision based on a feature or attribute. The value of a feature with the highest information gain or impurity reduction is used by the decision tree algorithm to recursively segment the data. The amount by which the target variable's uncertainty is reduced following the data's division based on a specific feature is measured by information gain. Reducing impurity results in more homogeneous subsets of data since impurity is a measure of the homogeneity of the target variable in a collection of data.

IV. EXPERIMENTAL RESULTS

From these findings, it is clear that even if the majority of studies use other algorithms, such as SVC and Decision trees, to identify patients with heart disease, KNN, Random Forest Classifier, and Logistic Regression produce a superior outcome to them. Our algorithms are faster and more precise than those employed by earlier studies. They also save a significant amount of money, making them very cost-effective. Additionally, the greatest accuracy of 88.5% achieved by KNN, Random Forest, and Logistic Regression is higher than or nearly equal to the accuracies attained in earlier studies that the larger number of medical attributes we used in the dataset we used has enhanced our accuracy.

We discover that the accuracy of the Random Forest is superior than other algorithms after using the machine learning technique for training and testing. With the help of the confusion matrix for each algorithm, accuracy is calculated. Here, the number of TP, TN, FP, and FN is provided. Using the equation for accuracy, the value has been calculated. It is concluded that extreme gradient boosting is the best with 98% accuracy. The comparison is shown below.

Algorithm	Accuracy
XG-boost	81.3%
SVM	80.2%
Logistic Regression	79.1%
Random Forest	98.0%
Naive Bayes	76.9%
Decision Tree	75.8%
Adaboost	73.6%

V. CONCLUSIONS

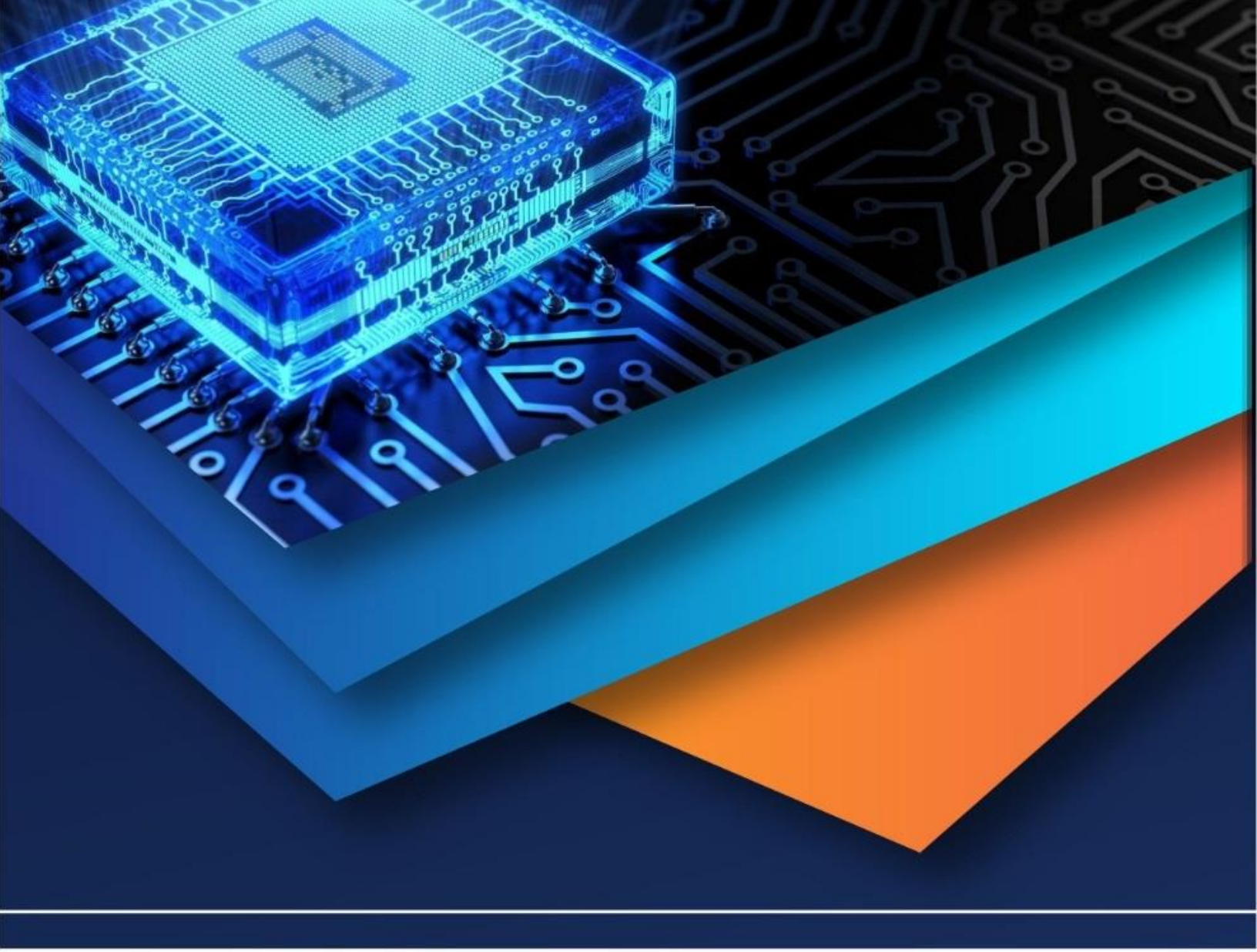
Based on symptoms, the system seeks to forecast heart disease. The system is set up to use symptoms as input and to provide results like predicting heart disease. The ML method has been successfully used to create the Heart Disease Predictor. The system seeks to provide ML algorithm analysis based on the accuracy parameter. In the future, wearable technology should be provided to identify additional indicators, such as blood pressure and pulse rate, in addition to eye blinking and yawning, to more effectively and correctly detect driver drowsiness and exhaustion and reduce the risk of traffic accidents.

VI. ACKNOWLEDGMENT

The preliminary project research paper on "Analysis of hybrid machine learning machine algorithm for heart disease prediction" is presented with great joy. I would want to take this opportunity to thank Mrs. Alpana Borse, my guide, for providing me with all the support and direction I required. I sincerely appreciate their thoughtful assistance. Their wise advice was quite beneficial. For his valuable guidance and suggestions, I am particularly grateful to Dr. Sonali Patil, Head of the Information technology Engineering Department at the Pimpri Chinchwad College Of Engineering, Akurdi, Pune. I would like to extend a particular thank you to Dr. Sonali Patil for offering several resources for our Project, including a laboratory with all necessary software platforms and ongoing guidance

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.
[2] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction Data Mining Knowledge. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.
[3] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, " Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Communication. , vol. 55, no. 1, pp. 54–61, Jan. 2017.
[4] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing," in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), Nov. 2016, pp. 184–189.
[5] Disease and symptoms Dataset –www.github.com.
[6] Heart disease Dataset-WWW.UCI Repository.com.
[7] Ajinkya Kunjir, Harshal Sawant, Nuzhat F. Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," in IEEE big data analytics and computational intelligence, Oct 2017 pp.23-25.
[8] Shanthi Mendis, Pekka Puska, Bo Norrving, World Health Organization (2011), Global Atlas on Cardiovascular Disease Prevention and Control, PP. 3– 18. ISBN 978-92-4-156437-3.
[9] Amin, S.U.; Agarwal, K.; Beg, R., "Genetic neural network based data mining in prediction of heart disease using risk factors", IEEE Conference on Information & Communication Technologies (ICT), 2013, vol., no.,pp.1227-31,11-12 April 2013.
[10] Palaniappan S, Awang R, "Intelligent heart disease prediction System using data mining techniques," IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2008., vol., no., pp.108115, March 31 2008-April 4 2008.
[11] B. Nithya, Dr. V. Ilango Professor, "Predictive Analytics in Health Care Using Machine Learning Tools and Techniques," International Conference on Intelligent Computing and Control Systems, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY